Airflow @Gojek
Streamlining Data Processing for Tableau Dashboards

Wanda Kinasih
BI @Gojek

Airflow Summit
Let’s flow together
September 19-21, 2023, Toronto, Canada
Hi! I’m Wanda!

Wanda Kinasih

- BI Analyst since 2016
- Now working as **BI Lead for Consumer Platform, Gojek**
- Experienced at:
  - SQL, Python
  - Data Visualisation using Tableau, Google Data Studio, Metabase
  - A/B Testing experiment
  - Google Cloud Project
  - Airflow, Pentaho
- Tableau Desktop Specialist Certified
Agenda

- Gojek Introduction
- Gojek Data Platform
- The Power of Airflow and Tableau Integration
A gojek group operating company
VISION

Become the Micro-Entrepreneurs Hero Brand.

MISSION

Create & scale up positive socio-economic impact on the ecosystem of users, driver partners, business & SMEs, as-well-as service providers.
Gojek aims to empower micro-entrepreneurs to make cities more accessible and engaging.

Tokopedia aims to democratise commerce through technology, empowering millions of consumers and merchants through its marketplace platform.

GoTo Financial is providing the comprehensive technological tools to help consumers and merchants thrive in the new economy.
OUR FOOTPRINT IN SOUTHEAST ASIA

Founded in Indonesia, Gojek now operates in **three** Southeast Asian countries.

**3 APPS:**
Consumer, Merchant Partner & Driver Partner

- Fulfils daily needs
- Increases turnover & business scale
- Optimizes the productivity of driver partners
2010
Gojek started commercial operations.

2015
Launched on-demand services app in Indonesia

2016
Launched GoPay

2021
Entered Vietnam & Singapore

2021
United with Tokopedia to create GoTo

The “go to” ecosystem for daily life combining on-demand e-commerce & financial tech services

THE JOURNEY SO FAR...
GoCar
Car ride-hailing service.

GoRide
Motorcycle taxi (ojek) ride-hailing service.

GoCar Protect+
Extra protection to feel safe on a trip.

GoBluebird
Bluebird taxi booking service.

GoTransit
Multi-modal journey planner solution.

GoCorp
Platform for corporate clients to easily access and monitor business-related trips for their employees.

GoMart
On-demand delivery from grocery and convenience stores.

GoFood
Food delivery service that provides consumers with convenient access to the best food options.

Cloud Kitchen
Shared kitchens for preparation of delivery-only meals.

GoSend
C2C product that provides consumers with fast and hassle-free instant and same-day delivery services.

GoSendAPI

GoBox
On-demand truck logistics service for large-sized deliveries.

GoShop
On-demand personal concierge service allowing consumers to shop for items and have them delivered within hours.
OUR IMPACT

Economy

Gojek contributed IDR 249 T to the national economy (equivalent to 1.6% of Indonesia’s GDP in 2020)

>1 million
GoFood merchant partners

>2.6 million
driver partners

Driver Partners

Gojek driver partners remain resilient during the pandemic.

Driver partners have experienced significant recovery through an increase in income of:

- 24% for GoRide partners
- 18% for GoRide partners

Compared to the beginning of the pandemic, drivers still have an income to support themselves and their families.

- 4 out of 5 partners feel the benefit from the time flexibility in their partnership with Gojek.
Gojek Data Platform

Data Source
- Google Cloud Platform
  - GO-DATA Production
    - GoPay
    - GoPoints
  - GO-DATA Production gofinance

Data Acquisition
- Google Cloud Platform
  - GO-DATA Production
    - Beagle
    - Sakaar

Data Governance
- Google Cloud Platform
  - GO-DATA Production
    - Predator
    - Data Discovery
    - Dexter
    - Hodor
    - Data Pier
    - Goender

Operations and Monitoring
- Google Cloud Platform
  - GO-DATA Production
    - Airflow
    - Stackdriver
    - Radar

Data Warehouse
- Google Cloud Platform
  - GO-DATA Production
    - BigQuery
    - Cloud Storage
    - Optimus
    - Digger

Data Visualization
- Google Cloud Platform
  - GO-DATA Production
    - Tableau
    - Mabase
(Simplified) Data Pipeline

Source (product)
- Kafka
- DB

Store (DWH)
- Data Lake
- Data Mart

Analyse (Analyst, Business)
- Tableau
Importance of Efficient Data Processing and Visualization

1. Informed Decision Making
   Efficient data processing and visualization enable organizations to quickly turn raw data into meaningful insights.

2. Faster Problem Solving
   By analyzing data in real-time and visualizing it in a comprehensible manner, organizations can identify issues early, troubleshoot efficiently, and minimize downtime.

3. Scalability and Performance
   Properly processed and visualized data allows systems to handle larger datasets without compromising speed or accuracy.

4. Data Quality Assurance
   Instantly detect inconsistencies, errors, and outliers, allowing data engineers to maintain high-quality datasets.
Gojek Tableau Dashboards

- >300 Data Sources
- >800 Dependencies
- >400 Daily Views
- >500 Active Users
- 55 Dashboard Creators
Lots of Data Sources in each Dashboard

Sample Dashboard

Owner: Wanda Kinasih
Modified: Sep 2, 2023, 7:52 PM

Edit Workbook

<table>
<thead>
<tr>
<th>Views</th>
<th>2</th>
<th>Data Sources</th>
<th>3</th>
<th>Custom Views</th>
<th>0</th>
<th>Subscriptions</th>
<th>0</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Actions</th>
<th>Connects to</th>
<th>Data comes from</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>bq.table_1</td>
<td></td>
<td>bq.table</td>
<td>Extract—Sep 27, 2021, 4:24 PM</td>
</tr>
<tr>
<td></td>
<td>bq.table_2</td>
<td></td>
<td>bq.table</td>
<td>Extract—Sep 27, 2021, 4:24 PM</td>
</tr>
<tr>
<td></td>
<td>bq.table_3</td>
<td></td>
<td>bq.table</td>
<td>Extract—Sep 27, 2021, 4:24 PM</td>
</tr>
</tbody>
</table>
Huge Data Sources in each Dashboard
Tableau Built-in Scheduler

- Only list of schedules
- Can’t set dependencies
- Can’t monitor each data source refresh easily
- Can’t see which job is failing
Integrating Tableau and Airflow

GCP Data Mart → Tableau Data Source Extracts → Tableau Dashboards → Tableau Viewers

Airflow Scheduling and Monitoring
How To Make Sure Each Data Source Wait For Their Dependencies?
Set Up Schedule Easily via Airflow

1. Upload dashboard and data sources in Tableau Server
2. Create yaml file for each data source
3. Monitor data sources update via Airflow
DAG Configuration (Py File)

```python
# list dependencies
run_wait bq_table_1 = ExternalTaskSensor(
    retries=1,
    retry_delay=timedelta(minutes=2),
    external_dag_id='d_1_dag1',
    external_task_id='bq_table_1',
    task_id='wait bq_table_1',
    execution_delta=timedelta(hours=3, minutes=30),
    dag=dag)

run_wait bq_table_2 = ExternalTaskSensor(
    retries=1,
    retry_delay=timedelta(minutes=2),
    external_dag_id='d_1_dag2',
    external_task_id='bq_table_2',
    task_id='wait bq_table_2',
    execution_delta=timedelta(hours=2, minutes=30),
    dag=dag)

run_wait bq_table_3 = ExternalTaskSensor(
    retries=1,
    retry_delay=timedelta(minutes=2),
    external_dag_id='d_1_dag3',
    external_task_id='bq_table_3',
    task_id='wait bq_table_3',
    execution_delta=timedelta(hours=3),
    dag=dag)

# refresh data source
run_refresh_tableau_data_source = DockerOperator(  
    task_id='refresh tableau_data_source',  
    command='"/opt/bi-tableau/config/folder/tableau_data_source.conf"',  
    # this consist configuration files for tableau refresh
    image='image.10/bi-tabcmd-app',  
    volumes=docker_volumes,  
    retries=5,  
    retry_delay=timedelta(minutes=3),  
    pool='tableau_refresh',  
    dag=dag)

# wait for dependencies
run_refresh_tableau_data_source.set_upstream(run_wait bq_table_1)
run_refresh_tableau_data_source.set_upstream(run_wait bq_table_2)
run_refresh_tableau_data_source.set_upstream(run_wait bq_table_3)
```

Dependencies in Google Cloud Bigquery

Tableau data source

Set upstream for each dependency
Simple Yaml File for Analysts and Business Users

version: 1
name: tableau.refresh.dataset_name.table_name
owner: email@gojek.com
schedule:
  start_date: "2022-01-01"
  interval: 0 22 ***

behavior:
  depends_on_past: false
  catch_up: false
  notify:
    - 'on': failure
      channels:
        - slack://#alert-slack

task:
  name: tableau
  config:
    ACTION: refresh_extract
    SERVER_URL: '{GLOBAL_TABLEAU_SERVER_URL}'
    SITE: site_name
    PROJECT: "Project Name"
    DATASOURCE: dataset_name.table_name

window:
  size: 24h
  offset: "0"
  truncate_to: h
labels:
  orchestrator: optimus

dependencies:
  - job: project_name.dataset_name.table_1
  - job: project_name.dataset_name.table_2
  - job: project_name.dataset_name.table_3

Dag name and schedule
Alerts
Tableau data source
Dependencies in Google Cloud BigQuery
Monitor Data Sources Easily via Airflow
Questions?

Let’s connect

https://www.linkedin.com/in/wandakinasih/